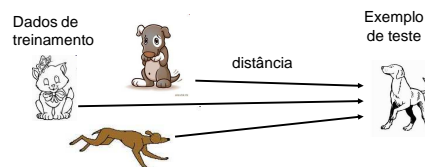
André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

## Métodos baseados em distância

- Consideram proximidade entre dados
  - Considera que dados similares tendem a estar em uma mesma região no espaço de entrada
- Aprendizado preguiçoso
  - Só olha os dados de treinamento quando precisa classificar novo objeto
- Exemplos:
  - Algoritmo k-vizinhos mais próximos
  - Raciocínio Baseado em Casos



## Similaridade x Dissimilaridade

- Similaridade
  - Mede o quanto dois objetos são parecidos
    - Quanto mais parecidos, maior o valor
  - Geralmente valor $\in$ [0, 1]
- Dissimilaridade
  - Mede o quanto dois objetos são diferentes
    - Quanto mais diferentes, maior o valor
  - Geralmente valor $\in$ [0, X]
- Medida de proximidade pode ser usada nos dois casos

## Distância Euclidiana

- Pode medir dissimilaridade de objetos com mais de um atributo
  - Para atributos com escalas de valores diferentes, pode ser necessário normalizar

$$dist = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$



## Distância de Minkowski

- Generalização da distância Euclidiana

$$dist = \left(\sum_{k=1}^{m} | p_k - q_k |^r\right)^{\frac{1}{r}}$$

- Valor de r leva a diferentes distâncias
  - 1 ($L_1$): Distância bloco cidade (Manhattan)
    - Hamming (valores binários)
  - 2 ($L_2$): Distância Euclidiana
  - ∞ ($L_\infty$): Distância suprema

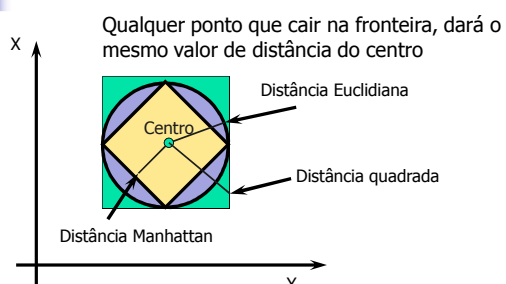


## Distância quadrada

- Simplificação da distância
  - Menor complexidade
  - Menor exatidão

$$dist = MAX(| p_k - q_k |)$$



## Relação entre distâncias

## Exercício

- Calcular a distância entre os exemplos abaixo usando as distâncias
  - Manhattan
  - Eucilidiana
  - Quadrada

Ex1 = (3, 1, 10, 2)

Ex2 = (2, 5, 3, 2)



## Exercício

- Encontrar a distância entre os exemplos abaixo utilizando a distância Manhattan
  - 110000, 111001, 000111, 001011, 100111, 101001

## Medidas de distâncias

- Têm, em geral, têm as propriedades:
  - Seja *d(p, q)* a distância (dissimilaridade) entre dois objetos *p* e *q*
    - $d(p, q) \geq 0 \; \forall \; p$ e $q$ e $d(p, q) = 0$ see $p = q$ (definida positiva)
    - $d(p, q) = d(q, p) \; \forall \; p$ e $q$ (simetria)
    - $d(p, r) \leq d(p, q) + d(q, r) \; \forall \; p, q$ e $r$ (desigualdade triangular)
- Medidas que satisfazem essas propriedades são denominadas métricas



## Medidas de similaridade

- Também têm propriedades bem definidas:
  - Seja *s(p, q)* a similaridade entre dois objetos *p* e *q*
    - $s(p, q) = 1$ (similaridade máxima) apenas se $p = q$
    - $s(p, q) = s(q, p) \; \forall \; p$ e $q$ (simetria)



## Dissimilaridade entre valores

- Sejam a e b dois valores de um atributo
  - Nominal $d(a,b) = \begin{cases} 1, \text{se } a \neq b \\ 0, \text{se } a = b \end{cases}$
    - s = 1 - d
  - Ordinal $d(a,b) = \frac{|a-b|}{n-1}$
    - s = 1 - d
  - Intervalar ou racional $d(a,b) = |a-b|$
    - s = - d ou s = 1/(1+d)



## Exercício

- Qual a distância entre os exemplos da tabela abaixo
- Usar distâncias
    - Euclidiana
    - Bloco cidade
    - Máxima

| Estado | Escolaridade | Altura | Salário | Classe |
|---|---|---|---|---|
| SP | Médio | 180 | 3000 | A |
| RJ | Superior | 174 | 7000 | B |
| RJ | Superior | 100 | 2000 | A |



## Similaridade entre vetores binários

- Frequentemente, objetos p e q têm apenas valores binários
- Similaridades podem ser computadas usando:
  - $M_{01}$ = número de atributos em que p = 0 e q = 1
  - $M_{10}$ = número de atributos em que p = 1 e q = 0
  - $M_{00}$ = número de atributos em que p = 0 e q = 0
  - $M_{11}$ = número de atributos em que p = 1 e q = 1



## Similaridade entre vetores binários

- Coeficiente de Casamento Simples

  CCS = num. de coinc. / num. de atributos
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- Coeficiente Jaccard

  J = num. coinc. 11 / num. Pelo menos.um $\neq$ 0
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

## Exercício

- Calcular disssimilaridade entre *p* e *q* usando coeficientes:
  - Casamento Simples
  - Jaccard

p = 1 0 0 1 1 0 1 0 1 1 1 0
q = 0 1 0 0 1 1 0 0 1 0 1 1



## Similaridade cosseno

- Muito usado para dados de textos
  - Grande número de atributos
  - Esparsos
- Sejam *p* e *q* vetores representando documentos
  - cos( *p, q*) = (*p* • *q*) / ||*p*|| ||*q*||
    - •: vector produto interno entre vetores
    - || *p* ||: é o tamanho (norma) do vetor *p*



## Classificação

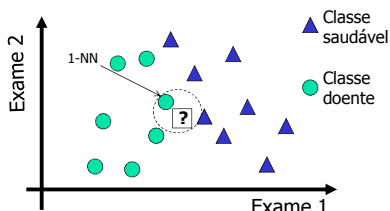
- Medidas de distância podem ser usadas para classificação de novos dados
  - Classificadores mais simples
    - K-NN
  - Dissimilaridade entre valores
  - Desempenho depende da medida de distância utilizada



## 1-vizinho mais próximo

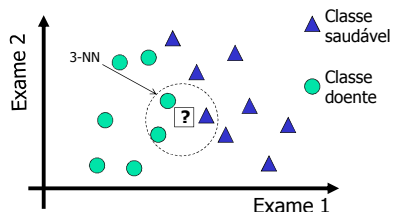
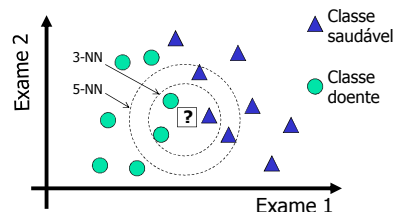
- Algoritmos *lazy* (preguiçoso)
  - Olha apenas os dados de treinamento quando precisa classificar novo objeto
  - Não constroem um modelo explicitamente
  - Diferente de classificadores *eager*, como SVMs e DTs
  - Baseados em informações locais
    - ADs, RNs e SVMs são baseados em informações globais



## 1-vizinho mais próximo





## Quantos vizinhos?

- K muito grande
  - Vizinhos podem ser muito diferentes
  - Predição tendenciosa para classe majoritária
  - Custo computacional mais elevado
- K muito pequeno
  - Não usa informação suficiente
  - Previsão pode ser instável
    - Ruído

Quantos vizinhos?
Exame 2
3-NN
?
Classe
saudável
Classe
doente
Exame 1
27/09/2012
André de Carvalho - ICMC/USP
25

Quantos vizinhos?
Exame 2
3-NN
5-NN
?
Classe
saudável
Classe
doente
Exame 1
27/09/2012
André de Carvalho - ICMC/USP
26

K-Vizinhos mais próximos
Seja k o número de vizinhos mais próximos a ser considerado
Para cada novo exemplo x
Definir a classe dos k exemplos mais próximos
Classificar x na classe majoritária entre seus vizinhos
27/09/2012
André de Carvalho - ICMC/USP
27

K-vizinhos mais próximos
Lento para classificar novos objetos
Seleção de atributos
Eliminação de objetos
Armazenar apenas protótipos das classes na memória
Algoritmos iterativos
Eliminação sequencial
Inserção sequencial
27/09/2012
André de Carvalho - ICMC/USP
28

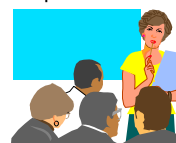
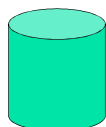
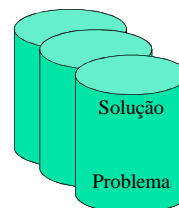
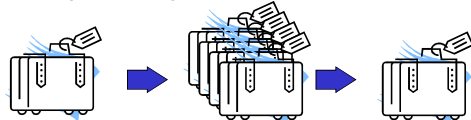
K-vizinhos mais próximos
Seleção de protótipos
Definir um protótipo por classe
Eliminação sequencial
Começa com todos os objetos
Descarta objetos corretamente classificados pelos protótipos
Inserção sequencial
Conjunto inicial vazio
Acrescenta objetos incorretamente classificados pelos protótipos
27/09/2012
André de Carvalho - ICMC/USP
29

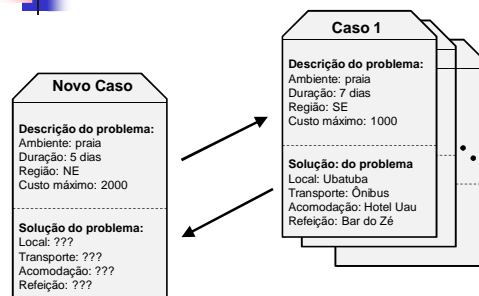
K-vizinhos mais próximos
Normalizar atributos
Ponderar atributos
Ponderar voto por distância entre exemplos
Regressão
Naturalmente incremental
27/09/2012
André de Carvalho - ICMC/USP
30

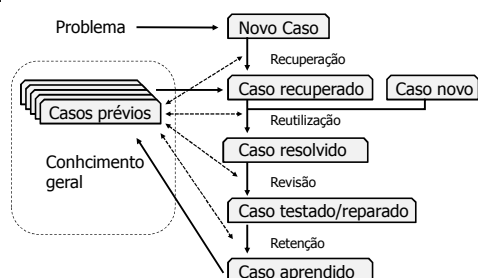
Exercício
Seja o seguinte cadastro de pacientes:
Nome Febre Enjôo Manchas Dores Diagnóstico
João sim sim pequenas sim doente
Pedro não não grandes não saudável
Maria sim sim pequenas não saudável
José sim não grandes sim doente
Ana sim não pequenas sim saudável
Leila não não grandes sim doente
27/09/2012 André de Carvalho - ICMC/USP 31

Exercício
Usar K-NN e os exemplos anteriores para definir as classes dos exemplos de teste
Usar k = 1, 3 e 5
Exemplos de teste
(Luis, não, não, pequenas, sim)
(Laura, sim, sim, grandes, sim)
27/09/2012 André de Carvalho - ICMC/USP 32

Exercício
Data a tabela abaixo, com k =1 e 3, definir a classe dos exemplos:
(RJ, Médio, 178, 2000)
(SP, Superior, 200, 800)
Estado Escolaridade Altura Salário Classe
SP Médio 180 3000 A
RJ Superior 174 7000 B
RS Médio 180 600 B
RJ Superior 100 2000 A
SP Fundam. 178 5000 A
RJ Fundam. 188 1800 A
27/09/2012 André de Carvalho - ICMC/USP 33

Raciocínio baseado em casos
Moda no passado: Sistemas Baseados em Regras
Dificuldade de especialistas em transformar experiência em regras
If ....
Then ...
Else...
EXPERIÊNCIA
REGRAS
27/09/2012 André de Carvalho - ICMC/USP 34

Raciocínio baseado em casos
EXPERIÊNCIA
BASE DE EXPERIÊNCIAS
Mas não uma BD!
27/09/2012 André de Carvalho - ICMC/USP 35

Como funciona RBC?
Resolve novos problemas adaptando soluções de problemas anteriores semelhantes
Solução
Problema
2
Nova solução
1
Novo problema
27/09/2012 André de Carvalho - ICMC/USP 36

Passos
Apresentar situação atual
Recuperar casos semelhantes da biblioteca
Adaptar solução
Que pacote de viagem comprar?
Casos semelhantes
Adaptação
27/09/2012
André de Carvalho - ICMC/USP
37

O que é um caso?
Existem dois tipos de casos
Casos de entrada:
Descrição de características de problemas específicos
Casos armazenados:
Casos anteriores
descrição, solução e resultados
27/09/2012
André de Carvalho - ICMC/USP
38

O que é um caso?
Um caso armazenado geralmente tem:
Uma parte caso
Sintomas
Usada para identificar o caso
Indexação e recuperação
Uma parte solução
Explica como este caso foi resolvido anteriormente de forma bem (mal) sucedida
Adaptada quando o caso é recuperado
27/09/2012
André de Carvalho - ICMC/USP
39

Raciocínio baseado em casos
Novo Caso
Descrição do problema:
Ambiente: praia
Duração: 5 dias
Região: NE
Custo máximo: 2000
Solução do problema:
Local: ???
Transporte: ???
Acomodação: ???
Refeição: ???
Caso 1
Descrição do problema:
Ambiente: praia
Duração: 7 dias
Região: SE
Custo máximo: 1000
Solução: do problema
Local: Ubatuba
Transporte: Ônibus
Acomodação: Hotel Uau
Refeição: Bar do Zé
27/09/2012
André de Carvalho - ICMC/USP
40

Ciclo de um sistema de RBC
Problema
Novo Caso
Recuperação
Caso recuperado
Caso novo
Casos prévios
Reutilização
Caso resolvido
Conhcimento geral
Revisão
Caso testado/reparado
Retenção
Caso aprendido
27/09/2012
41

Conclusão
Aprendizado baseado em distância
Conceitos básicos
KNN
Raciocínio Baseado em Casos
Exemplos
27/09/2012
André de Carvalho - ICMC/USP
42

Perguntas
27/09/2012
André de Carvalho - ICMC/USP
43